DOCUMENT RESUME

ED 120 240 TM 005 199

AUTHOR Dinero, Thomas E.; Haertel, Edward
TITLE A Computer Simulation Investigating the Applicability of the Rasch Model with Varying Item Discriminations.
PUB DATE 20 Apr 76
NOTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, California, April 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Comparative Analysis; *Computer Programs; Goodness of Fit; Individual Differences; *Item Analysis; *Mathematical Models; Matrices; Probability; Scoring; *Simulation; Test Construction
IDENTIFIERS Item Calibration (Tests); Item Discrimination (Tests); *Rasch Model

ABSTRACT

This paper will discuss the results of a series of computer simulations comparing the Rasch logistic model to a series of models departing to various degrees from its assumption of equal discrimination power for all items. The results have implications for test construction and test scoring, indicating how closely the conventional raw score corresponds to the mathematically correct score, in which each item response is weighted by an index of the item discrimination. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED120240

A COMPUTER SIMULATION INVESTIGATING THE APPLICABILITY OF THE RASCH MODEL WITH VARYING ITEM DISCRIMINATIONS

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Thomas E. Dinero
Kent State University

Edward Haertel
Chicago Public Schools

National Council on Measurement in Education
April 20, 1976

TM005 199

A Computer Simulation Investigating the Applicability of the Rasch Model with Varying Item Discriminations

In the classical model of item analysis, two principle characteristics of an item merit attention--these are, of course, the item difficulty and item discrimination. In many situations, these indices seem to offer the test user important, non-redundant information about his test. Most champions of the classical model would be careful to admonish the user to be sensitive to the interdependency of his results and the subjects who have yielded them.

In 1960, however, Rasch (Probablistic Models for Some Intelligence and Attainment Tests, cited in Lord, F. and Novick, M., 1968; Whitely and Dawis, 1974; Wright and Panchapakesan, 1969) presented three models to explain misreadings, number of words read, and general achievement; each of these is a two parameter model, encorporating only the ability of the person and the difficulty of the measurement to explain the observed data. The most impressive implication of the models is that item calibration and individual measurement are independent of both each other and the situation in which they take place.

The suggestion that an examinee's item score depends on only his ability and the difficulty of the item is an inherently pleasing one to many people. Without test artifacts like item discrimination to get in the way, the individual is pitted clearly against his criterion, and would thus, one might expect,

supply us with neatly interpretable data. Whether the picture is as clear as this has yet, of course, to be shown.

The present research artificially generated the results of several hypothetical tests for which the effects of item discriminations varied. Fit to Rasch's assumptions was predicted on the fact that his third model may be understood as a two-parameter logistic function. With this bridge to more general models, then, the Rasch assumption of equal discriminability could be tested.

The clearest demonstration of the relationship between the person and the item is the item characteristic curve. Here one theorizes the latent ability of the person (or class of people) plotted against the probability of getting a particular item correct. At least since Guilford (1936), it has been assumed that, within the ability range of the test, this probability is best described by the normal ogive function. However, the assumption of normality has been seen by some to be delimiting. By positing the logistic test model,

$\psi(x) = \frac{e^x}{1+e^x}$ one can neatly avoid the restriction since Haley (cited in Birnbaum, 1968), has shown that if $\Phi(x)$ is the cumulative normal distribution function $|\Phi(x) - \psi[(1.7)x]| < 0.01$ for all X. Within the context of test theory this model takes on a specific form credited to Birnbaum, $P_g(\theta) = \psi[\beta(\theta-\alpha)]$, where α is the difficulty of the item, β is the item discrimination, and θ the examinee ability. The probability density function of this model is:

$$f_g(U_g=1|\theta) = \frac{e^{\beta(\theta-\alpha)}}{1+e^{\beta(\theta-\alpha)}} \quad \text{and}$$

$$f_g(U_g=0|\theta) = \frac{e^{0}}{1+e^{\beta(\theta-\alpha)}} = \frac{1}{1+e^{\beta(\theta-\alpha)}}$$

where U_g is 1 if the examinee responds correctly on item g and U_g is 0 if he does not (Birnbaum, 1968). This, then, is the most general statement of the logistic model encorporating maximum information about the item and the examinee.

Rasch (1966 a,b) has presented a model which can be seen as a simplification of this, one in which $X=\beta(\theta-\alpha)$ can be explained in terms of θ and α alone, the item discrimination (β) having been assumed constant across items (hence, here equal to one). The implications of this lie in the fact that one can estimate θ independently of α and vice versa. As Wright and Panchapakasan (1965) have indicated, however, the model implies that:

1. the model is unidimensional;
2. there are no strong relationships among persons or items other than those specified by the model so that responses of persons to items are stochastically <u>independent</u> given their parameters in the model;
3. items and persons do not differ substantially with respect to other response factors not represented in the model such as item discrimination, person sensitivity, guessing, or indifference. (p. 2)

The author added that since few can write items as a predetermined discriminating level, it is most feasible to discard "grossly dissimilar items (p. 4)," resulting in a set of

items with "similar discrimination and minimal guessing". If one were to have a decision rule for doing this, he would then be assured a fortiori of building a test in conformity with the Rasch model.

The present discussion describes one solution to the establishment of such a criterion. A Monte Carlo computer program using the Rasch model was designed to input person and item parameters, generate probabilities of success, simulate a test-taking situation, produce the raw item score matrix, and estimate the parameters of the Rasch item characteristic curve. All four subsections may be used independently of each other, parameters can be read in or generated internally, and link-ups with other subsections are determined only by the intent of the user. The subsection which estimates the parameters of the Rasch item characteristic curve will accept as input either a raw item score matrix or a matrix of probabilities of success. In addition, the data-generating function follows Birnbaum's (1968) three parameter model, and the data calibration follows Wright and Panchapakesan (1969); this allowed the present methodology: generation of data using Birnbaum for simulation, and analysis of this data using Wright and Panchapakesan is calibration based on Rasch's model. Poor calibration would then suggest lack of robustness of the Rasch calibration to departures from homogeneity of item discrimination.

The Simulation Program

There are three general foci in the present FORTRAN simulation. The first reads item difficulties, discriminations,

or person abilities, or generates them according to user specifications. Following this, the parameters are combined according to the Birnbaum formulation into a person x item matrix of probabilities. In the second (and actual simulation) phase, a series of random numbers is generated, each number being between 0 and 1; these numbers are compared with the probabilities generated in phase one and the "raw data" matrix is generated according to the rule:

$$a_i = 1 \text{ if } P(a_i=1) > \text{random number}$$

$$a_i = 0 \text{ if } P(a_i=1) < \text{random number}$$

The matrix of a_i's could have been read in at this point instead of being generated.

The third phase involves item calibration based on either the matrix of raw item scores or the person X item matrix of probabilities generated in the first phase of the simulation. Rasch (1966 a, b) has shown that, assuming the double parameter model, total unweighted scores, that is $\sum_i a_{in}$ for person (or score group) n are sufficient statistics for latent ability, which is estimated by the person (or score group) parameter θ. Wright and Panchapakesan (1969) have elaborated Rasch's original least squares approach; in addition, they have presented a maximum likelihood estimation which is more precise.

Several points need to be made about this estimation. First, there is one and only ability level for any one score (or score group). Second, item calibration (that is, determining the alpha or item difficulties) generally precedes person

measurement (determining the thetas or person abilities). Third, if $\sum_{j=1}^{n} a_{ij} = n$, that is, if any item is gotten correctly by all n people, it is useless for calibration. Similarly, if $\sum_{i=1}^{k} a_{in} = k$, i.e., all items were gotten correctly by person j, that person's ability cannot be estimated, and his responses contain no information concerning the relative difficulties of the items.

The least squares and maximum likelihood methods are briefly treated here; the discussion follows Wright and Panchapakesan closely. The estimation of the item difficulty is based upon the assumption that, within any score group, the probability of success on item i is approximately the proportion of people within that score group who produced a correct response to that item. With the estimate of the difficulty scaled so that the mean difficulty equals zero, the standard error of estimate is derived from the variances of these probabilities using the assumption that the actual responses to a given item within a given score group are binomially distributed. Estimation of person measurement is exactly parallel to this.

The maximum likelihood estimates are necessary only for item calibration; the item estimates, generated first, can be used to calculate directly person abilities. Initially, the implicit equations for item difficulties and person abilities are solved simultaneously, using an iterative procedure. Once the items have been calibrated, the ability estimate for any examinee depends upon nothing but his total raw score. More-

over, any set of calibrated items may be combined to form a new test, and a similar set of implicit equations may be solved iteratively to determine the ability estimate corresponding to any possible raw score on the new test. Additionally, these estimates of theta and alpha are used to calculate the standard error of estimate of the estimated difficulty.

For each item its goodness-of-fit to the Rasch model is computed by forming a standard deviate

$$Y_{ji} = \frac{a_{ji} - E(a_{ji})}{V(a_{ji})^{1/2}}$$

where a_{ji} is the obtained item score for person j on item i, $E(a_{ji})$ is the estimate of a_{ji} based on item difficulties (α_i) and person ability (θ_j), and $V(a_{ji})^{1/2}$ is the standard deviation of the a_{ji}. The squares of the standard deviates summed over people yield an approximate X^2 with N - 1 degrees of freedom which can be used to test the fit of item i to the model.

Procedure

The central concern of the present research was the effect of item discriminations on fit to the Rasch model. It was believed that a certain tolerance is allowed in the application of the theory. Exactly how much, of course, was not known. Degree of fit would be based on the degree to which item discriminations were the same, that is, did not vary among themselves. This degree of fit was therefore operationalized as the variance of the item discriminations. For the present simulations, variances were assumed to be .05, .10, .15, .20, and .25. One run was also made at $\sigma^2 = 0$ to indicate

degree of accuracy of the item calibration. As there was also some question about the shape of the distribution of these values, this quality was also varied. Three forms were used, normal, uniform, and positively skewed. This latter form is thought to be the most reasonable for a well-constructed test since discrimination values should never be negative; with a mean of one, the distribution would skew right. The actual shape was operationalized as approximately a chi-square distribution with one degree of freedom. Since the Rasch procedure automatically scales the person and item parameters in such a way as to make the average item discrimination equal to one, this value was taken as the mean of all distributions studied. There were thus sixteen simulation runs, one for the pure Rasch model and three at each degree of discrimination variability; all had means equal to one. All parameters except item discrimination were held constant. For each run a test length of 30 items was employed, with item difficulties randomly sampled from a normal distribution with mean 0 and standard deviation 1. The obtained random sample which was used for all runs, had a mean of .113 and a sample standard deviation of .940, the range being from -1.553 to 2.070.

For each of the sixteen simulation runs, two calibrations were performed. First, the person X item matrix of probabilities was calibrated. This approximates the result of administering the test to an infinite sample and calibrating the data obtained in the conventional manner. The only difference between this calibration and calibration on an infinite sample lies in the

fact that when the a priori probability matrix is calibrated directly, the parameters may vary continuously, while in calibrating actual data the ability estimates take on a set of discrete values, corresponding to each possible total raw score between 1 and k-1 items correct on a k-item test.

The second calibration of each of the sixteen simulation runs was performed on a data matrix obtained by simulating an administration of the test to 75 persons and analyzing the obtained raw data matrix. The abilities of these persons were sampled at fixed intervals from a normal distribution with mean 0 and variance 1.5. The obtained sample had a mean of 0.00 and a sample variance of 1.475. While it is known that the best item calibration is done with a good deal of replicability within each score group, hence large N (Whitely and Dawis, 1974), the computer time and cost were prohibitive for this. The n of 75 was considered sufficient because (1) the additional calibration was obtained on an "infinite" sample and (2) the 75 "persons" used were "centered on the test", i.e., the test was of exactly the right difficulty for them, resulting in a very efficient administration with respect to amount of information obtained.

In the computer model, all simulations allow the item difficulties and discrminations and the person abilities either to be read in or generated internally. For the present reseach all three were generated randomly with the following characteristics. For all runs, the item difficulties were the same, having been randomly selected form the unit normal distribution. The

person abilities were also normally distributed about a mean of zero, except they had a variance of four. With these data fixed, sixteen simulation runs were attempted. The first used a standard default option built into the program and generated a unit vector of item discriminations; this, then, was the run where the precision of the item calibration routine could be tested since the input was purely Rasch-conforming. Each of the remaining simulations, however, deviated from the Rasch assumption of similar discriminations in two ways. Five of the runs had discriminations uniformly distributed with mean equal to one and variance equal respectively to .05, .10, .15, .20, and .25, each run showing increasingly stronger deviation from Rasch's assumption. For each of these runs, discriminations were sampled at fixed intervals from the appropriate uniform distribution. The next five simulations had discriminations normally distributed around a mean of one and variances respectively .05, .10, .15, .20, and .25. Values were once more sampled at fixed intervals.

The remaining five analyses were based on the chi-square distribution with one degree of freedom. This distribution has a mean of one and a variance of two. Data points were selected in the following manner. Since thirty item discriminations were needed, the chi-square PDF was broken into thirty equal areas; the mean of each area constituted the preliminary data point. These thirty points, with their mean of one and variance of two, were then converted to a data set having a mean of one and a variance of .25 using a linear transforma-

tion. This set of points was adjusted slightly to obtain the desired range of discriminations while holding the first two moments constant, and, finally, the obtained set was linearly transfored to each of a set of points having a mean of one and the variances used above (.05, .10, .15, .20, .25).

RESULTS

The results indicated that the Rasch calibration procedure is robust to departures from homogeneity of item discrimination but that any tendency for this robustness to be lost does conform to Rasch's assumption. Table 1 presents what are believed to be the salient characteristics of the calibration upon which one might focus. Both item fit and person fit are described for P matrix calibration and new data matrix calibration. The criteria of interest are the mean fit and the most extreme point of lack of fit. While they are both self-explanatory, it is felt that the latter deserves some explication. The extreme instance of misfit may be misinterpreted unless it is borne in mind that 1) it is a single score and by its nature an extreme one and 2) for many of the simulations, they are belived to be outliers. As a last point, the standard errors of estimate of person and item parameters are rarely less than .1 logit and can be quite a bit greater.

Several patterns were noted in the results, and, while there have been no statistical tests to confirm them, they have been included here.

First, as is quite striking in Table 1, the average item fit for both the P matrix and the raw data matrix was identically zero (including rounding errors). In contrast, the

fit of persons was zero or negligible for the fits with zero beta variance or variance of .05 of the distribution were normal or skewed.

Second, the poorest fit seemed to be for the uniform distribution, where both the average and maximum misfits were quite a bit larger than for either of the other distributions.

Third, the patterns acorss increasing variability are clearer in the theoretical (P matrix) calibration; as can be expected, the random error introduced in the simulation of test-taking clouded the issue. Here, both the uniform and normal distributions of betas showed the expected pattern: the fit of item and people became worse as the variance of the betas increased. The pattern for the skewed distribution was not so clear.

SUPPLEMENTARY ANALYSIS

An argument for the use of item discrimination may still be made in terms of the extraction of maximal information from the test. In order to assess the degree to which unweighted total score approximates the mathematically correct scoring in which each response is weighted by its item's discrimination, the mathematically correct scoring was correlated with the unweighted number of items correct for each simulation. The minimum of these correlations across all sixteen simulating runs was .8069. The magnitude of these correlations suggests

that whatever loss of information the use of unweighted raw scores might entail could be compensated for by a slight increase in test length. This conclusion, unfortunately, cannot be generalized to the case where a test is of inappropriate difficulty for the examinees. In this case, a correlation may be induced between item difficulty and item discrimination, because items at one end of the continuum of item difficulties represented in the test will function better, and hence appear more discriminating, than items at the other end of the difficulty continuum.

CONCLUSIONS

The present research suggests that the lack of an item discrimination parameter in the Rasch model does not result in poor calibration in the presence of varying item discriminations. While the robustness of the model to other departures from assumptions remains to be investigated, such studies are also indicated for the normal ogive model, more general logistic models, etc. Until such time as it is shown to be either inadequate or inferior to some other model, the use of the simplist model is to be recommended, if only on the basis of mathematical elegance and the sufficiency of total number of items correct as a statistic for subject ability.

The substitution of equal item discriminations, rather than maximum item discriminations as a goal in item writing, appears counter-intuitive to the test construction expert steeped in classical test theory. While it is true that

a highly discriminating item is capable of providing more information concerning the placement of an individual on the continuum of some latent trait, the highly discriminating item functions over a narrower range of abilities than a less discriminating item. An item with perfect discrimination would provide complete information about a single point on the ability continuum and no information about any other point. Therefore, for any given test, there will exist an optimal range of discrimination. If the test characteristic curve is to rise steeply through a narrow range of abilities, more highly discriminating items will be desirable than if the test is to function over a broad range of abilities.

No guidelines can be provided indicating a specific range of item discriminations which may be tolerated, first, because the model is highly robust to differing discriminations and, second, because in the actual application of the model the true values of the discriminations are unknwon. Item discriminations are estimated following calibration, by regressing probability of success of ability in the (linear) logistic matric. The wrose the fit of an item, the larger the standard error of estimate of its discrimination may be. In the light of these considerations, the authors suggest Wright's (1969) approximate X^2 statistics for the evaluation of fit.

TABLE 1

Degree of Misfit of Item Calibration For All

			Fit of Items				Fit of Persons			
			P Matrix		Data Matrix*		P Matrix		Data Matrix*	
Run	σ^2_β	Dist.	Mean Misfit	Maximum Misfit	Mean Misfit	Maximum Misfit	Mean Misfit	Maximum Misfit	Mean Misfit	Maximum Misfit
1	.00	-	0	- .1127	0	- .5113	.0000	.1132	.0600	-1.2834
2	.05	U	0	-2.8300	0	-3.3306	-.0406	3.0608	-.0893	1.7688
3	.10	U	0	-3.0390	0	-3.3803	-.1203	2.7481	-.1363	2.1622
4	.15	U	0	-3.1833	0	-3.5172	-.1696	3.0961	-.2436	-2.6026
5	.20	U	0	-3.2942	0	-3.7293	-.2058	3.0978	-.2277	1.4506
6	.25	U	0	3.3843	0	-4.3912	-.2358	3.0981	-.3614	-2.5156
7	.05	N	0	.1453	0	.6251	-.0027	.1310	-.0075	.6004
8	.10	N	0	.2203	0	.8546	-.0049	.2628	-.0772	.7857
9	.15	N	0	.2827	0	.7216	-.0071	.3882	-.0607	.8206
10	.20	N	0	.3387	0	1.0069	-.0092	.5054	-.0543	.8114

*computed by score group

TABLE 1 (continued)

Degree of Misfit of Item Calibration for All

Run	σ^2_β	Dist.	Fit of Items				Fit of Persons			
			P Matrix		Data Matrix*		P Matrix		Data Matrix*	
			Mean Misfit	Maximum Misfit	Mean Misfit	Maximum Misfit	Mean Misfit	Maximum Misfit	Mean Misfit	Maximum Misfit
12	.05	S	0	.3011	0	.4901	-.0060	.1016	-.0560	1.0583
13	.10	S	0	.5438	0	1.1696	-.0287	.1139	-.0656	1.5914
14	.15	S	0	.4856	0	.9101	.0237	.3542	.0543	1.7754
15	.20	S	0	.3889	0	.6427	.0247	.4361	.0336	.3533
16	.25	S	0	.6251	0	1.0491	-.0246	.3434	-.0451	.6491

TABLE 2

Correlations Between the Unweighted Total Score Approximations and the Weighted Total Scores for All Simulations

Simulation	σ^2_β	Distribution	$r_{x \cdot x(c)}$
1	.00	--	.9847
2	.05	U	.8123
3	.10	U	.8195
4	.15	U	.8069
5	.20	U	.8547
6	.25		.8183
7	.05	N	.9921
8	.10	N	.9875
9	.15	N	.9877
10	.20	N	.9885
11	.25	N	.9887
12	.05	S	.9888
13	.10	S	.9851
14	.15	S	.9827
15	.20	S	.9872
16	.25	S	.9786

Limitations of the Present Research and Suggestions for Future Research

In this study, the only source of misfit which was introduced into the data was nonhomogeneity of item discrimination. The calibration procedure proved quite robust to perturbations of this kind. Actual data, however, are influenced by a wide variety of effects, e.g., guessing, carelessness when items are too easy, practice effects which distract the shape of the item characteristic curve and/or induce violations of the assumption of local independence of persons and items.

These additional sources of misfit raise several questions:

1. If more than one parameter is to be estimated for each item, is discrimination the best choice to accompany difficulty, or would more variance be accounted for by a parameter representing, say, level of asymptote of the item characteristic curve (sensitivity to guessing)?

2. Would the Rasch calibration procedure be less robust to variation in item discrimination if those variations occured in the context of other sources of misfit?

3. If variations in item discrimination alone do not preclude the use of the Rasch model, what evidence is there that the normal ogive model is superior to the Rasch model in fitting actual data?

REFERENCES

Anderson, E. S. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-39

Anderson, J., Kearney, G. E., and Everett, A. B. An evaluation of Rasch's structural model for test items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability, Part V of F. M. Lord and M. R. Novick Statistical Theory of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Guilford, J. P. Psychometric Methods, New York: McGraw-Hill, 1936.

Lord, F. M. and Novick, M. R. Statistical Theories of Mental Test Scores, Reading, Massachusetts: Addison-Wesley Publishing Company, 1938.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57. (a)

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.) Readings in Mathematical Social Science. Chicago: Science Research Associates, 1966, 89-108. (b)

Whitely, S. E. and Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 3, 163-178.

Wright, B. and Panchapakesan, W. A. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.